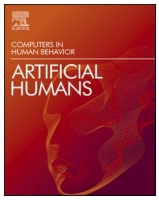
Computers in Human Behavior: Artificial Humans 2 (2024) 100070

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



On trust in humans and trust in artificial intelligence: A study with samples from Singapore and Germany extending recent research

Christian Montag [a,*], Benjamin Becker [b,c], Benjamin J. Li [d]

[a] Department of Molecular Psychology, Institute of Psychology and Education, Ulm University, Ulm, Germany
[b] State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, China
[c] Department of Psychology, The University of Hong Kong, China
[d] Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Artificial intelligence
Trust
Humans
Attitudes towards artificial intelligence
Personality

ABSTRACT

The AI revolution is shaping societies around the world. People interact daily with a growing number of products and services that feature AI integration. Without doubt rapid developments in AI will bring positive outcomes, but also challenges. In this realm it is important to understand if people trust this omni-use technology, because trust represents an essential prerequisite (to be willing) to use AI products and this in turn likely has an impact on how much AI will be embraced by national economies with consequences for the local work forces. To shed more light on trusting AI, the present work aims to understand how much the variables *trust in AI* and *trust in humans* overlap. This is important to understand, because much is already known about trust in humans, and if the concepts overlap, much of our understanding of trust in humans might be transferable to trusting AI. In samples from Singapore (n = 535) and Germany (n = 954) we could observe varying degrees of positive relations between the *trust in AI/humans* variables. Whereas *trust in AI/humans* showed a small positive association in Germany, there was a moderate positive association in Singapore. Further, this paper revisits associations between individual differences in the Big Five of Personality and general attitudes towards AI including trust.

The present work shows that *trust in humans* and *trust in AI* share only small amounts of variance, but this depends on culture (varying here from about 4 to 11 percent of shared variance). Future research should further investigate such associations but by also considering assessments of trust in specific AI-empowered-products and AI-empowered-services, where things might be different.

1. Introduction

Artificial intelligence (AI) is in-built in a growing number of products/services people use daily. The current changes can be conceptualized as the beginning of what the (co-)founder of DeepMind Mustafa Suleyman describes as the coming wave (Suleyman & Bhaskar, 2023), whereas societies around the world likely will see disruptions due to the AI-technology. It already becomes apparent that economies will be shaped by the AI technology, but likely with varying effects (Furman & Seamans, 2019). AI's impact on societies is hard to forecast because AI represents an omni-use technology and many variables must be considered to understand how AI impacts upon people's attitudes toward this technology, the willingness to use it and then of course what this means for societies. To illustrate this a bit further: A recent framework summed up with the acronym IMPACT the idea that an **I**nterplay of the variables **M**odality, **P**erson, **A**rea, **C**ulture/**C**ountry and **T**ransparency will be relevant to understand how attitudes towards AI and well-being when interacting with AI-products/services are shaped (Montag, Ali, Al-Thani, & Hall, 2024; Montag, Nakov, & Ali, 2024). Again, such attitude formation will also impact national economies because more positive attitudes towards AI will likely also result in more use of AI technologies (in line with this idea see that more positive attitudes towards AI are also associated with more trust in ChatGPT; Montag & Ali, 2023). Hence, the IMPACT framework postulates that variables such as personality of the AI-users, the area where AI is operating, regulation efforts of countries and explainable AI will all need to be considered to grasp AI's impact on users and societies.

We believe that a prerequisite for a successful transit from pre-AI- to AI-societies will be that people trust AI. In this regard, the question arises if more *trust in humans* goes also along with more *trust in AI*. In

* Corresponding author. Department of Molecular Psychology, Institute of Psychology and Education, Ulm University, Helmholtzstr. 8/1, 89081 Ulm, Germany.
E-mail address: mail@christianmontag.de (C. Montag).

https://doi.org/10.1016/j.chbah.2024.100070
Received 30 November 2023; Received in revised form 7 May 2024; Accepted 7 May 2024
Available online 10 May 2024

other words, are people who report to trust humans also persons who more trust AI? This question is in so far highly relevant because abundant research exists both on a psychological, neuroscientific, and behavioral economic level shedding light on the underpinnings of trust in humans (Alós-Ferrer & Farolfi, 2019; Riedl & Javor, 2012). If one would observe that the concepts *trust in humans* and *trust in AI* are highly overlapping one might be able to transfer some of the knowledge from the concept *trust in humans* to the concept *trust in AI*. This is not necessarily the case though, because a recent work showed on a questionnaire level that the concepts *trust in humans* and *trust in AI* were unrelated (Montag, Klugah-Brown, et al., 2023). Further, in this work only individual differences in *trust in humans* were significantly associated with individual differences in brain structure, specifically in striato-thalamic and prefrontal regions which play a role in several behavioral domains, including motivational approach-avoidance and social cognitive processes (Feng et al., 2021; Xu et al., 2023). However, the initial study had several limitations, and we explicitly mentioned the preliminary character of the findings (p. 7), also because of the "highly exploratory approach" (p. 6). In detail, in the former study by Montag, Klugah--Brown, et al. (2023) "only" a small group of 90 male Chinese participants could be studied (but see also that MRI data was investigated which restricts the opportunities to include large sample sizes). Further, in this work the questionnaires assessing *trust in humans* and *trust in AI* did not align perfectly to each other, because *trust in humans* was assessed via several items from a personality inventory via a five answer format (NEO-PI-R; Costa & McCrae, 2008) and a single item was used to assess *trust in AI* from the ATAI with an eleven-answer-format (Sindermann et al., 2021). Of note, the answer format was chosen in this study to stick with the original answer format of the scales. Some of the shortcomings of this recent study are aimed to be overcome in the present study and this is outlined in the following.

First, we recruited two large samples with different cultural background (German sample, Singapore sample). This approach can be seen as highly relevant, because much research in the field of the psychological sciences is hampered by the WEIRD problem (Henrich, Heine, & Norenzayan, 2010). Hence, study participants often have a **W**estern, **E**ducated, **I**ndustrialized, **R**ich and **D**emocratic background and it is not clear if observations can be transferred to other populations. Against the background of the WEIRD problem, this time we aim to understand if what has been observed in a Chinese sample will be transferable to both a Central European (Germany) and South-East Asian sample (Singapore). Second, in the recent work only male Chinese participants could be recruited (Montag, Klugah-Brown, et al., 2023), and as a consequence, we strive for much more diverse samples, here (we aimed at a gender balanced ratio at each site). Third, we aimed to maximize the chance to find an association between the concept of *trust in humans* and *trust in AI* by administering near identical items which only differ in the words humans vs. AI (I trust humans vs. I trust AI). In addition, these items are responded to with the same answer format. In case, when following this approach, still only very small associations would be observable, this would further speak for independent constructs.

We hypothesized based on our initial work that *trust in humans* and *trust in AI* would not be associated (or if, only very small positive associations would be visible). Please note that this work including the hypotheses were preregistered (https://osf.io/v58yg). This work is of course not a replication of the study by Montag, Klugah-Brown, et al. (2023), but should be seen as an extension with the changes of several elements in the study design.

In the context of the preregistered hypotheses, we also mention that we investigated links between personality and attitudes towards AI again. In line with the IMPACT framework (Montag, Nakov, & Ali, 2024), we expected that personality would be linked to individual differences in AI-attitudes. In particular, we expected that higher neuroticism would be linked to more fear of AI, as this was observed in a Chinese and German sample earlier (Sindermann et al., 2022). Further minor hypotheses will be addressed in the result section for reasons of completeness.

2. Methods

2.1. Background on the recruitment process

We recruited two samples for the present study to revisit the question on how *trust in humans* and *trust in AI* would correlate with each other - but making the changes in the study design as mentioned above. In the following, information on each sample is presented including the data cleaning process to ensure good data quality at each site. The samples recruited stem from Singapore and Germany. All final analyzed data in this work is shared with the community allowing further investigation (see section 2.5).

2.2. On the sample from Singapore

An initial sample of N = 803 participants was recruited through the Centre for Information Integrity and the Internet (IN-cube) research institute in Singapore. Participants were recruited from the general population, needed to be at minimum 21 years old and for the present study purpose filled in the ATAI questionnaire (assessing **A**ttitudes **T**owards **A**rtificial **I**ntelligence; Sindermann et al., 2021), a further *trust in humans* item and the short BFI-10 scale (Rammstedt & John, 2007) to assess individual differences in the Big Five of Personality. From the initial sample of N = 803 participants, n = 193 failed an attention check, which resulted in a sample of n = 610 participants. Age information was appropriate, and no one needed to be filtered out in this regard. The same was true for the gender information. As the BFI consists of several inverse coded items, we filtered out those with 9 or 10 same answers on the ten items. After this cleaning step n = 584 participants remained. In the remaining cases, we did further data checks regarding monotonous answer behavior: The ATAI consists of five items assessing acceptance and fearing AI (two vs. three items). We filtered out all participants who had the same answers on all five items (this strategy is in particular debatable regarding same answers in the middle (hence 3), because technically you could have everywhere a neutral response). For reasons of consistency in data cleaning we decided on this strategy at all sites. The final sample consisted of n = 535 participants (256 males, 279 females; mean-age: 43,96, SD = 12,47, age-range: 22–76 years). The study was approved by the local ethics committee at Nanyang Technological University, Singapore.

2.3. On the sample from Germany

N = 1151 participants could be recruited via the Bilendi Research Institute in Cologne, Germany. These participants needed to be at minimum 18 years old and were recruited from the general population. From this initial sample n = 54 were deleted, because they did not fill in the survey in one piece (they stopped somewhere and came back later). Four participants were excluded, because they reported the third gender (this group is unfortunately too small to run analysis on). Further participants were excluded, when they failed an attention check, did not report full consent and did not fulfill language criteria. This all resulted in n = 1082 participants. Data cleaning of the BFI data (answering 9 or 10 items of 10 BFI items with the same answers) led to a reduction of the sample to n = 1053. Further data cleaning of BFI and ATAI with missing data led to a sample of n = 968. A final data cleaning step comprised excluding participants with same answers on all five ATAI items. This led to a final sample of n = 954 participants (471 males, 483 females; mean-age: 44,66, SD = 14,33, age-range: 18–76). The study was approved by the local ethics committee at Ulm University, Ulm, Germany.

2.4. Questionnaires

In the present work the Attitudes toward Artificial Intelligence (ATAI) scale was administered (Sindermann et al., 2021). The ATAI scale consists of five items, which yield two dimensions: accepting AI (two items) and fearing AI (three items). The accepting AI facet also includes the single item called "I trust artificial intelligence.", which is used for further analysis. In the present work items were answered via a five-answer-format ranging from strongly disagree (1) to strongly agree (5). Internal consistencies were as follows for *accepting AI* in the ATAI scale: $\alpha = 0.65$ (Singapore); $\alpha = 0.77$ (Germany). Internal consistencies were as follows for *fearing AI* in the ATAI scale: $\alpha = 0.76$ (Singapore); $\alpha = 0.76$ (Germany). The internal consistencies need to be seen in light of the very few items applied. To be able to contrast *trust in AI* with *trust in humans* also the item called "I trust humans." was administered. To make the items directly comparable in terms of the answer format, the same five-answer-format was rolled out (see a contrasting approach with different answer formats in Montag, Klugah-Brown, et al., 2023). We also administered an item asking about trustworthiness ("I am trustworthy."), to be able to investigate same personality correlations as in an older work (Evans & Revelle, 2008).

In Germany, the German version of the ATAI scale was administered, in the Singapore sample the English version was administered. Both language versions can be found in Sindermann et al. (2021). In addition to the ATAI scale and the aligned "trust in humans" item, we also assessed in the present study the Big Five of Personality with the brief BFI-10. The German and English versions were used as presented in Rammstedt and John (2007).

2.5. Statistical analyses

We investigated how trust in humans and trust in AI would correlate in each sample. We present Pearson correlations here. As we also have included personality data and the full ATAI measure, we present in the correlation tables all associations of interest (see Tables 2 and 3). Further descriptive statistics are presented (see Table 1) and also *trust in humans/trust in AI* levels are contrasted across the samples by means of MANCOVAs (see results for further information on this approach). For the statistical analysis both SPSS (29.0.1.0) and the Jamovi package (2.4.8.0) were used. Fig. 1 depicting the scatterplot figure was created with the *scatr* plugin in Jamovi. Fig. 2 was created with Apple's Keynote (13.2). Data can be found at the OSF (https://osf.io/fg62k/).

3. Results

Descriptive statistics for the two samples are presented in Table 1. Following up on our hypotheses from the preregistration, we investigated the relationship between *trust in humans* and *trust in AI* variables. As one can see in the scatterplot in Fig. 1 the Singapore sample showed a moderate positive association between these variables ($r = 0.335$, $p < 0.001$), the German sample showed a small positive association between these variables ($r = 0.192$, $p < 0.001$). As mentioned in the preregistration we also assessed in Singapore and in Germany the additional items "I trust artificial intelligence easily." vs. "I trust humans easily." (German: Ich vertraue künstlicher Intelligenz/Menschen leichtfertig). Interestingly, this wording indicating some impulsiveness/risk-taking, led to a substantial increase of overlap in the German and Singapore samples (Germany: $r = 0.446$, $p < 0.001$; Singapore: $r = 0.440$, $p < 0.001$). Finally, we mentioned in the preregistration to also test if the sample from Singapore would be characterized by higher acceptance of AI (the trust item is part of this scale). This turned out to be true ($t_{(1250,996)=}6.941$, $p < 0.001$, Cohen's $d = 0.359$). Interestingly, the Singapore sample also scored higher on the fear of AI scale than the German sample (this was not expected; $t_{(1487)=}7.829$, $p < 0.001$, Cohen's $d = 0.423$).

Aside from this, in all samples more *trust in humans* compared to *trust in AI* was reported. See Fig. 2. MANCOVA with trusting humans/trusting AI as dependent variables (and gender/country being independent variables, age as a covariate) revealed a significant "influence" (no causality implied) of country on trusting artificial intelligence ($F_{(1,1484)} = 45.136$, $p < 0.001$, $eta^2 = 0.030$), but not on trusting humans. Gender was significantly associated with the trust in artificial intelligence variables ($F_{(1,1484)} = 14.203$, $p < 0.001$, $eta^2 = 0.009$; males > females), but not with the trust in human variable. No gender by country interaction effect was significant and therefore this is not further followed up. Please see that - against the significant age associations reported in Tables 2 and 3 - we inserted age as a covariate in the MANCOVA analysis.

Paired t-tests revealed that in all samples *trust in humans* was significantly higher than in *trust in AI* (Singapore: $t_{(534)} = -4.14$, $p < 0.001$, Cohen's $d = -0.179$; Germany: $t_{(953)} = -12.15$, $p < 0.001$, Cohen's $d = -0.393$). This effect was most pronounced in the German sample. Further, the descriptive statistics in Table 1 reveal that in Singapore higher trust in AI was reported compared to the German sample. Regarding trust in humans the descriptive statistics revealed that the Singapore and German sample had about same expressions.

A further research question revisited associations between personality and the ATAI scale in the Singapore and German sample. Here we could observe that higher neuroticism, lower agreeableness and lower

Table 1
Descriptive statistics.

	Country	N	Missing	Mean	Median	SD	Minimum	Maximum
I trust AI.	Singapore	535	0	3.20	3	0.946	1	5
	Germany	954	0	2.84	3.00	1.008	1	5
I trust humans.	Singapore	535	0	3.40	4	0.989	1	5
	Germany	954	0	3.34	3.00	0.971	1	5
Fearing AI	Singapore	535	0	9.22	9	2.638	3	15
	Germany	954	0	8.09	8.00	2.694	3	15
Accepting AI	Singapore	535	0	6.84	7	1.521	2	10
	Germany	954	0	6.23	6.00	1.773	2	10
Openness	Singapore	535	0	3.11	3.00	0.656	1.00	5.00
	Germany	954	0	3.43	3.50	1.045	1.00	5.00
Conscientiousness	Singapore	535	0	3.48	3.50	0.782	1.50	5.00
	Germany	954	0	3.79	4.00	0.856	1.00	5.00
Extraversion	Singapore	535	0	2.73	3.00	0.869	1.00	5.00
	Germany	954	0	3.03	3.00	1.065	1.00	5.00
Agreeableness	Singapore	535	0	3.59	3.50	0.779	1.00	5.00
	Germany	954	0	3.12	3.00	0.842	1.00	5.00
Neuroticism	Singapore	535	0	2.82	3.00	0.859	1.00	5.00
	Germany	954	0	2.67	2.50	1.012	1.00	5.00

Table 2
With correlations from Singapore.

		Fearing AI	Accepting AI	I trust AI.	I trust humans.	I am trustworthy.	O	C	E	A	N	Age
Fearing AI	Pearson's r	–										
	df	–										
	p-value	–										
	N	–										
Accepting AI	Pearson's r	−0.231***	–									
	df	533	–									
	p-value	<0.001	–									
	N	535	–									
I trust AI.	Pearson's r	−0.109*	0.882***	–								
	df	533	533	–								
	p-value	0.012	<0.001	–								
	N	535	535	–								
I trust humans.	Pearson's r	0.014	0.322***	0.335***	–							
	df	533	533	533	–							
	p-value	0.748	<0.001	<0.001	–							
	N	535	535	535	–							
I am trustworthy.	Pearson's r	−0.069	0.229***	0.183***	0.339***	–						
	df	533	533	533	533	–						
	p-value	0.110	<0.001	<0.001	<0.001	–						
	N	535	535	535	535	–						
Openness	Pearson's r	0.005	0.001	0.041	−0.053	0.037	–					
	df	533	533	533	533	533	–					
	p-value	0.914	0.978	0.340	0.217	0.391	–					
	N	535	535	535	535	535	–					
Conscientiousness	Pearson's r	−0.248***	0.081	0.056	0.125**	0.344***	0.086*	–				
	df	533	533	533	533	533	533	–				
	p-value	<0.001	0.061	0.198	0.004	<0.001	0.046	–				
	N	535	535	535	535	535	535	–				
Extraversion	Pearson's r	−0.056	0.138**	0.144***	0.190***	0.106*	0.074	0.202***	–			
	df	533	533	533	533	533	533	533	–			
	p-value	0.192	0.001	<0.001	<0.001	0.014	0.087	<0.001	–			
	N	535	535	535	535	535	535	535	–			
Agreeableness	Pearson's r	−0.268***	0.110*	0.099*	0.164***	0.317***	−0.042	0.381***	0.107*	–		
	df	533	533	533	533	533	533	533	533	–		
	p-value	<0.001	0.011	0.022	<0.001	<0.001	0.335	<0.001	0.013	–		
	N	535	535	535	535	535	535	535	535	–		
Neuroticism	Pearson's r	0.250***	−0.201***	−0.188***	−0.145***	−0.178***	0.016	−0.354***	−0.377***	−0.306***	–	
	df	533	533	533	533	533	533	533	533	533	–	
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001	0.718	<0.001	<0.001	<0.001	–	
	N	535	535	535	535	535	535	535	535	535	–	
Age	Pearson's r	−0.177***	−0.105*	−0.106*	0.030	0.137**	−0.003	0.314***	0.045	0.250***	−0.271***	–
	df	533	533	533	533	533	533	533	533	533	533	–
	p-value	<0.001	0.015	0.015	0.489	0.002	0.948	<0.001	0.296	<0.001	<0.001	–
	N	535	535	535	535	535	535	535	535	535	535	–

Note. *p < 0.05, **p < 0.01, ***p < 0.001.

Table 3
With correlations from Germany.

		Fearing AI	Accepting AI	I trust AI.	I trust humans.	I am trustworthy.	O	C	E	A	N	Age
Fearing AI	Pearson's r	–										
	df	–										
	p-value	–										
	N	–										
Accepting AI	Pearson's r	−0.587***	–									
	df	952	–									
	p-value	<0.001	–									
	N	954	–									
I trust AI.	Pearson's r	−0.488***	0.908***	–								
	df	952	952	–								
	p-value	<0.001	<0.001	–								
	N	954	954	–								
I trust humans.	Pearson's r	−0.106**	0.186***	0.192***	–							
	df	952	952	952	–							
	p-value	0.001	<0.001	<0.001	–							
	N	954	954	954	–							
I am trustworthy.	Pearson's r	−0.100**	0.146***	0.121***	0.348***	–						
	df	952	952	952	952	–						
	p-value	0.002	<0.001	<0.001	<0.001	–						
	N	954	954	954	954	–						
Openness	Pearson's r	−0.018	0.059	0.056	0.160***	0.105**	–					
	df	952	952	952	952	952	–					
	p-value	0.580	0.067	0.081	<0.001	0.001	–					
	N	954	954	954	954	954	–					
Conscientiousness	Pearson's r	−0.107***	0.071*	0.051	0.046	0.231***	0.141***	–				
	df	952	952	952	952	952	952	–				
	p-value	<0.001	0.029	0.116	0.151	<0.001	<0.001	–				
	N	954	954	954	954	954	954	–				
Extraversion	Pearson's r	−0.071*	0.128***	0.134***	0.274***	0.132***	0.227***	0.149***	–			
	df	952	952	952	952	952	952	952	–			
	p-value	0.027	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	–			
	N	954	954	954	954	954	954	954	–			
Agreeableness	Pearson's r	−0.136***	0.219***	0.214***	0.530***	0.261***	0.125***	0.092**	0.154	***	—	
	df	952	952	952	952	952	952	952	952	–		
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.004	<0.001	–		
	N	954	954	954	954	954	954	954	954	–		
Neuroticism	Pearson's r	0.231***	−0.109***	−0.095**	−0.191***	−0.130***	−0.055	−0.254***	−0.244***	−0.221***	–	
	df	952	952	952	952	952	952	952	952	952	–	
	p-value	<0.001	<0.001	0.003	<0.001	<0.001	0.092	<0.001	<0.001	<0.001	–	
	N	954	954	954	954	954	954	954	954	954	–	
Age	Pearson's r	−0.016	−0.136***	−0.128***	−0.011	−0.002	−0.009	0.085**	−0.049	0.026	−0.244***	–
	df	952	952	952	952	952	952	952	952	952	952	–
	p-value	0.617	<0.001	<0.001	0.736	0.950	0.770	0.008	0.132	0.423	<0.001	–
	N	954	954	954	954	954	954	954	954	954	954	–

Note. *p < 0.05, **p < 0.01, ***p < 0.001.

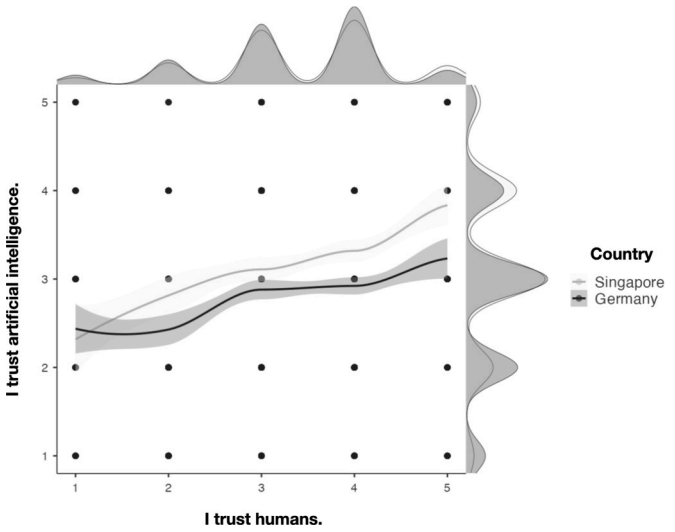


Fig. 1. Associations between trusting humans and trusting artificial intelligence reveal different associations with varying effect sizes for the different samples investigated.

conscientiousness were associated with higher fearing AI (with very mild to mild effect sizes). As we mentioned in the preregistration also to look at trust in AI – personality associations (item 2 of the ATAI), we report among others a positive association with extraversion and a negative association with neuroticism with mild effect sizes (this was not expected). In addition, lower neuroticism, more agreeableness and more extraversion were associated with more accepting AI (again very mild to mild effect sizes). As a further quality check of the data, we also report associations between the Big Five of Personality and individual differences in self-reported trusting humans and trustworthiness. Higher extraversion, higher agreeableness and lower neuroticism was associated with higher trust in humans. Higher conscientiousness and higher agreeableness were in particular linked to persons reporting to be more trustworthy (mild effect sizes). This is in line with earlier observations with more lengthy questionnaires (Evans & Revelle, 2008) and underlines the quality of the data of the present work (this was also preregistered).

4. Discussions

The present work built on a recent observation combining self-report data and structural brain imaging to understand links between trust in humans and trust in artificial intelligence (Montag, Klugah-Brown, et al., 2023). This previous study was limited by only investigating a comparably small male Chinese sample. Further, to see if associations might in parts also rely on the questionnaires employed with different answer formats, we aligned the wording and the answer format in this work, so that same sounding items were used to assess trust in humans/AI, just differing in the targets to be trusted. In our preregistration basing on our initial findings, we expected none or only a small positive association. This hypothesis was only in parts confirmed. In the German sample the positive correlation is around 0.20, which would fall in the realm of a small effect size (but clearly in the earlier work by Montag et al. using a different trust scale plus an 11-answer format on the side of the ATAI null correlations were observed). This said, $r = 0.20$ means just about 4% shared variance between variables, hence showing that these variables overlap only in very small parts. And one should not forget that the study design here maximized chances to find overlap between the variables. In more contrast to our hypothesis is the data from Singapore, where a moderate association around >0.30 could be observed. Although samples might differ in other variables not assessed (mean-age and gender ratio is mostly comparable), we believe that culture could play a role to explain differences in the observed association strength between *trust in humans* and *trust in AI* in both samples. This mirrors also in the investigated contrasts of the levels of *trust in humans* or *trust in AI* in each respective sample. Whereas in all samples *trust in humans* was higher than *trust in AI*, the effect was most pronounced in Germany and less pronounced in Singapore (also due to higher trust in AI in Singapore compared to Germany). Further complicating the interpretation of the present findings are observations showing that trust in humans might be interpreted in different ways across cultures. Research suggests that although items asking participants if most people can be trusted indeed seem to encompass also persons not belonging to one's inner circle, but the size of the radius might differ (for instance narrower in Confucian societies; Delhey, Newton, & Welzel, 2011). This needs to be considered in future research (and also see that our *trust in humans* item was formulated in a different way).

Going beyond the findings from the initial study (Montag, Klugah-Brown, et al., 2023), we observe that *trust in humans* and *trust in AI* seem still to be (rather) distinguishable concepts, even when wording of the items and the answer format are aligned, which makes it much easier to see overlap between the concepts. But it is also true that it is hard to decipher the exact effect size of the associations (ranging here from about 0.192 to 0.335), because in line with the recently proposed IMPACT model (Montag et al., 2024), Country/Culture represents a variable likely influencing the attitudes towards AI (including the here assessed trust item belonging to the ATAI scale). And again, the true associations might be well beyond what we see in this work, because the study design made it likely to observe such associations (see also

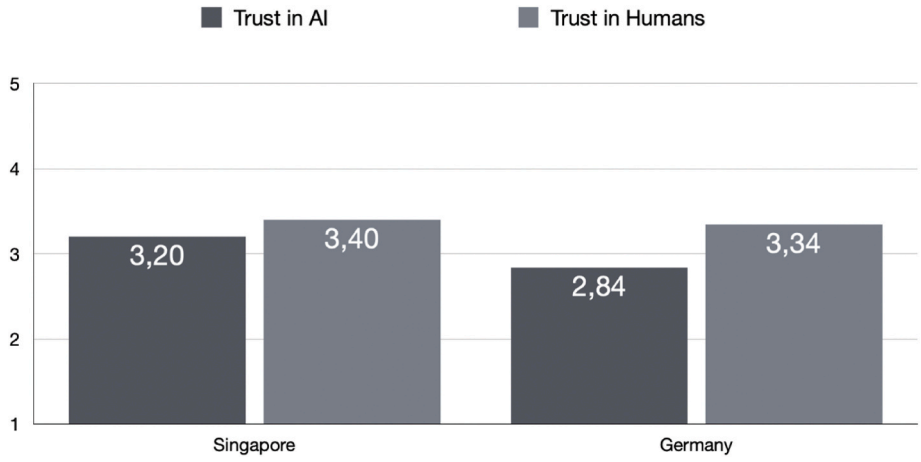


Fig. 2. Contrasting levels of trust in AI and trust in humans in the two samples under investigation.

limitations).

For reasons of completeness, and also as a quality check of the present data, we revisited personality associations with the ATAI. In line with an earlier work on this topic (Sindermann et al., 2022), we observed higher neuroticism to be linked to higher fearing AI. Hence, persons tending to experience in general more negative emotions also report more fearing AI, which is not surprising. Beyond, this also new associations appeared such as that also higher conscientiousness was associated with lower fearing AI (of note the Sindermann et al. study differed regarding the BFI assessment used and the answer format of the ATAI). This result might be explainable by the fact that more conscientious people might be more informed about the coming AI wave and therefore might be more prepared in how to handle unforeseen situations arising from the AI revolution. Beyond this, we further observed that higher agreeableness was also linked to both more acceptance of AI (and lower fearing of AI). Perhaps this has to do with the more cooperative and empathic nature of the agreeable person to approve/reject such statements (Melchers et al., 2016). The personality associations reported in this discussion were mostly in the mild effect size area. In so far, many other variables likely play a role to shape attitudes towards AI (including trust), and such variables have been mentioned in the introduced IMPACT framework describing the **I**nterplay of **M**odality, **P**erson, **A**ffect, **C**ountry/**C**ulture and **T**ransparency categories. The present empirical evidence could be seen as support that both the ***P-*** and the ***C-***category indeed are relevant.

The present study comes with several limitations. The study is of self-report nature, whereas participants might have answered in a socially desirable fashion, or they were lacking introspection. Beyond this, the present study is of cross-sectional nature, therefore no causality can be established between the variables. Further, the items were presented together in rather short surveys. It would be interesting to understand how the links between *trust in humans* and *trust in AI* would be, if these items are answered on different days or in different situations. In other words, the present study made it likely due to the setup to find overlap and the real effect sizes might be below what has been observed here. Moreover, one can criticize that AI is a very broad concept and people associate very different things with this concept (and see also the discussed cultural issues concerning the understanding of the trust term). Therefore, it is important to also investigate associations between *trust in distinct AI-products* and *trust in humans,* in the near future. This said, recent research suggests that general trust in AI items are indeed linked to trusting very different AI products (Montag & Ali, 2023; Montag, Klugah-Brown, et al., 2023; Sindermann et al., 2021) – so a very general formulated AI trust item has already some diagnostic value. One could also criticize to assess the trust in AI/trust in humans constructs with single items as in the present study. Here, this was done because we had the chance to get the items included in two panels collecting data in the general population. It would be important to run more nuanced trust in humans/AI measures in the future. Finally, we mention that wording of the items in this context seems to play a huge role. In the result section one could see that just adding the word “easily” to both trust items results in a boost of effect size regarding the associations (perhaps due to the perceived riskiness facet of the now presented item).

Concluding, the present study observes that *trust in humans* and *trust in AI* are (rather) separate constructs, but it is also true that these correlations differed in effect size from small to the lower moderate effect size area in the different samples. Therefore, more research is needed (including also neuroscientific work) to further understand if *trust in humans* and *trust in AI* are related constructs, which would also mean that perhaps some of the knowledge about trust formation in humans could be transferred to the area of trusting AI. Against the limitations and the early stage of the research field also the present work should be seen as preliminary.

CRediT authorship contribution statement

Christian Montag: Writing – original draft, Methodology, Data curation, Conceptualization. **Benjamin Becker:** Writing – review & editing. **Benjamin J. Li:** Writing – review & editing, Data curation, Conceptualization.

Declaration of competing interest

Dr. Montag reports no conflict of interest. However, for reasons of transparency Dr. Montag mentions that he has received (to Ulm University and earlier University of Bonn) grants from agencies such as the German Research Foundation (DFG). Dr. Montag has performed grant reviews for several agencies; has edited journal sections and articles; has given academic lectures in clinical or scientific venues or companies; and has generated books or book chapters for publishers of mental health texts. For some of these activities he received royalties, but never from gaming or social media companies. Dr. Montag mentions that he was part of a discussion circle (Digitalität und Verantwortung: https://about.fb.com/de/news/h/gespraechskreis-digitalitaet-und-verantwortung/) debating ethical questions linked to social media, digitalization and society/democracy at Facebook. In this context, he received no salary for his activities. Finally, he mentions that he currently functions as independent scientist on the scientific advisory board of the Nymphenburg group (Munich, Germany). This activity is financially compensated. Moreover, he is on the scientific advisory board of Applied Cognition (Redwood City, CA, USA), an activity which is also compensated. The other authors also not report a conflict of interest.

Acknowledgements

Any opinions, findings, conclusions or recommendations expressed in this publication do not reflect the views of the Government of the Hong Kong Special Administrative Region or the Innovation and Technology Commission.

This study was partially supported by the Centre for Information Integrity and the Internet (IN-cube), Nanyang Technological University, Singapore.

References

Alós-Ferrer, C., & Farolfi, F. (2019). Trust games and beyond. *Frontiers in Neuroscience, 13*, 887. https://doi.org/10.3389/fnins.2019.00887

Costa, P. T., Jr., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment, Vol. 2. Personality measurement and testing* (pp. 179–198). Sage Publications, Inc.. https://doi.org/10.4135/9781849200479.n9

Delhey, J., Newton, K., & Welzel, C. (2011). How general is trust in “most people”? Solving the radius of trust problem. *American Sociological Review, 76*, 786–807. https://doi.org/10.1177/0003122411420817

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality, 42*(6), 1585–1593. https://doi.org/10.1016/j.jrp.2008.07.011

Feng, C., Eickhoff, S. B., Li, T., Wang, L., Becker, B., Camilleri, J. A., et al. (2021). Common brain networks underlying human social interactions: Evidence from large-scale neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews, 126*, 289–303. https://doi.org/10.1016/j.neubiorev.2021.03.025

Furman, J., & Seamans, R. (2019). AI and the economy. *Innovation Policy and the Economy, 19*, 161–191. https://doi.org/10.1086/699936

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302). https://doi.org/10.1038/466029a. Article 7302.

Melchers, M. C., Li, M., Haas, B. W., Reuter, M., Bischoff, L., & Montag, C. (2016). Similar personality patterns are associated with empathy in four different countries. *Frontiers in Psychology, 7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00290.

Montag, C., & Ali, R. (2023). *Can we assess attitudes toward AI with single items? Associations with existing attitudes toward AI measures and trust in ChatGPT in two German speaking samples.* ResearchSquare. https://doi.org/10.21203/rs.3.rs-3325511/v1

Montag, C., Ali, R., Al-Thani, D., & Hall, B. J. (2024). On artificial intelligence and global mental health. *Asian Journal of Psychiatry, 91*, Article 103855. https://doi.org/10.1016/j.ajp.2023.103855

Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., et al. (2023). Trust toward humans and trust toward artificial intelligence are not associated: Initial

insights from self-report and neurostructural brain imaging. *Personality Neuroscience, 6*, e3. https://doi.org/10.1017/pen.2022.5

Montag, C., Nakov, P., & Ali, R. (2024). Considering the IMPACT framework to understand the AI-well-being-complex from an interdisciplinary perspective. *Telematics and Informatics Reports, 13*, Article 100112. https://doi.org/10.1016/j.teler.2023.100112

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big five inventory in English and German. *Journal of Research in Personality, 41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Riedl, R., & Javor, A. (2012). The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics, 5*(2), 63–91. https://doi.org/10.1037/a0026318

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., et al. (2021). Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English language. *KI - Künstliche Intelligenz, 35*(1), 109–118. https://doi.org/10.1007/s13218-020-00689-0

Sindermann, C., Yang, H., Elhai, J. D., Yang, S., Quan, L., Li, M., et al. (2022). Acceptance and fear of artificial intelligence: Associations with personality in a German and a Chinese sample. *Discover Psychology, 2*(1), 8. https://doi.org/10.1007/s44202-022-00020-y

Suleyman, M., & Bhaskar, M. (2023). *The coming wave: Technology, power, and the twenty-first century's greatest dilemma*. Crown.

Xu, T., Zhou, X., Kanen, J. W., Wang, L., Li, J., Chen, Z., et al. (2023). Angiotensin blockade enhances motivational reward learning via enhancing striatal prediction error signaling and frontostriatal communication. *Molecular Psychiatry, 28*(4). https://doi.org/10.1038/s41380-023-02001-6. Article 4.